Appendix C. Source and Reliability of the Estimates

SOURCE OF DATA

Most of the estimates in this report are based on data obtained in October 1986 from the Current Population Survey (CPS) conducted by the Bureau of the Census and from supplementary questions to the CPS. Some estimates are based on data obtained from the CPS in earlier years. The monthly CPS deals mainly with labor force data for the civilian noninstitutional population. Questions relating to labor force participation are asked about each member in every sample household. In addition, supplementary questions regarding school enrollment are asked about all eligible household members 3 years old and over each October.

The present CPS sample was initially selected from the 1980 census files with coverage in all 50 States and the District of Columbia. The sample is continually updated to reflect new construction. Each month, approximately 59,500 housing units are eligible for interview. Of this number, about 2,500 units are visited, but interviews are not obtained because the occupants are not found at home after repeated calls or are unavailable for some other reason.

The following table provides a description of some aspects of the CPS sample designs in use during the referenced data collection periods.

Description of the October Current Population Survey

Time period	Number of sample areas	Housing units eligible	
		Interviewed	Not interviewed
October 1985 to 1986	729	57,000	2,500
October 1984	[1]629/729	59,000	2,500
October 1981 to 1983	629	58,000	2,500
October 1980	629	63,000	3,000
October 1978 to 1979	614	53,500	2,500
October 1972 to 1977	461	45,000	2,000
October 1971	449	45,000	2,000
October 1967 to 1970	449	48,000	2,000
October 1965 to 1966	357	33,500	1,500
October 1960	333	33,500	1,500
October 1955	230	21,000	500-1,000
October 1947 to 1950	68	21,000	500-1,000

[1]The CPS was redesigned following the 1980 Decennial Census of Population and Housing. During phase-in of the new design, housing units from both designs were in the sample.

ESTIMATION

The estimation procedure used in this survey involves the inflation of the weighted sample results to independent estimates of the total civilian noninstitutional population of the United States by age, race, sex, and Hispanic categories. These independent estimates are based on statistics from the decennial censuses of population; statistics on births, deaths, immigration and emigration; and statistics on the strength of the Armed Forces. The independent population estimates used in this report to obtain data for 1981 and later are based on the 1980 decennial census. In earlier reports in this series (P-20), data for 1972 through 1980 were obtained using independent population estimates based on the 1970 decennial census. Estimates for earlier years were based on earlier censuses.

RELIABILITY OF THE ESTIMATES

Since the CPS estimates were based on a sample, they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaires, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey: sampling and nonsampling. The accuracy of a survey result depends on both types of errors, but the full extent of the nonsampling error is unknown. Consequently, particular care should be exercised in the interpretation of figures based on a relatively small number of cases or on small differences between estimates. The standard errors provided for the CPS estimates primarily indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in responses and enumeration, but do not measure any systematic biases in the data. (Bias is the difference, averaged over all possible samples, between the sample estimates and the desired value.)

Nonsampling variability. Nonsampling errors can be attributed to many sources; e.g., inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information, inability to recall information, errors made in data collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, and failure to represent all units with the sample (undercoverage).

Undercoverage in the CPS results from missed housing units and missed persons within sample households. Overall undercoverage as compared to the level of the 1980 decennial census is about 7 percent. It is known that CPS undercoverage varies with age, sex, and race. Generally, undercoverage is larger for males than for females and larger for Blacks and other races combined than for Whites. Ratio estimation to independent age-sex-race population controls, as described previously, partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that missed persons in missed households or missed persons in interviewed households have different characteristics from those of interviewed persons in the same age-sex-race group. Further, the independent population controls used have not been adjusted for undercoverage in the 1980 census.

For additional information on nonsampling error including the possible impact on CPS data when known, refer to Statistical Policy Working Paper 3, *An Error Profile: Employment as Measured by the Current Population Survey*, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, 1978 and Technical Paper 40, *The Current Population Survey: Design and Methodology*, Bureau of the Census, U.S. Department of Commerce.

Sampling variability. The standard errors given in tables C-1 through C-4 are primarily measures of sampling variability, that is, of the variation that occurred by chance because a sample rather than the entire population was surveyed. The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Some statements in the report may contain estimates followed by a number in parentheses. For those statements one has only to add to and subtract from the estimate the number in parentheses to calculate upper and lower bounds of the 90-percent confidence interval. For example, if a statement contains the phrase "grew by 1.7 percent (± 1.0)", the 90-percent confidence interval for the estimate, 1.7 percent, would be from 0.7 percent to 2.7 percent.

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common type of hypothesis appearing in this report is that the population parameters are different. An example of this would be comparing the percent of adults who were high school graduates in 1986 to those in 1976. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the parameters are different when, in fact, they are identical.

All statements of comparison in the text have passed a hypothesis test at the 0.10 level of significance. This means that, for all the differences cited in the text, the estimated difference between characteristics is greater than 1.6 times the standard error of the difference.

Comparability of data. Caution should be used when comparing estimates for 1981 and later, which reflect 1980 census-based population controls, with estimates from earlier years. This change in population controls had relatively little impact on summary measures such as means, medians, and percent distributions, but did have a significant impact on levels. For example, use of 1980-based population controls results in about a 2-percent increase in the civilian noninstitutional population and in the number of families and households. Thus, estimates of levels for 1981 and later will differ from those for earlier years by more than what could be attributed to actual changes in the population, and these differences could be disproportionately greater for certain subpopulation groups than for the total population.

In addition, the estimates in this report for 1985 and 1986 are based on revised survey weighting procedures for persons of Hispanic origin. In previous years the estimation procedures used in this survey involved the inflation of weighted sample results to independent estimates of the noninstitutional population by age, sex, and race. There was, therefore, no specific control of the survey estimates for the Hispanic population. During the last several years, the Bureau of the Census has developed independent population controls for the Hispanic population by sex and detailed age groups and has adopted revised weighting procedures to incorporate these new controls. It should be noted that the independent population estimates include some, but not all, illegal immigrants.

Note when using small estimates. Summary measures (such as medians, and percent distributions) are shown in the report only when the base is 75,000 or greater. Because of the large standard errors involved, there is little chance that summary measures would reveal useful information when computed on a smaller base. Estimated numbers are shown, however, even though the relative standard errors of these numbers are larger than those for corresponding percentages. These smaller estimates are provided primarily to permit such combinations of the categories as serve each data user's needs. Also, care must be taken in the interpretation of small differences. For instance, even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Standard error tables and their use. In order to derive standard errors that would be applicable to a large number of estimates and could be prepared at a moderate cost, a number of approximations were required. Therefore, instead of providing an individual standard error for each estimate, generalized sets of standard errors are provided for various types of characteristics. As a result, the sets of standard errors provided give an indication of the order of magnitude of the standard error of an estimate rather than the precise standard error.

The figures presented in tables C-1 through C-4 are approximations to the standard errors of various school enrollment estimates for persons in the United States. To obtain the approximate standard error for a specific characteristic, the appropriate standard error in tables C-1 through C-4 must be multiplied by the factor for that characteristic given in table C-5. These factors must be applied to the generalized standard errors in order to adjust for the combined effect of the sample design and the estimating procedure on the value of the characteristic. Standard errors for intermediate values not shown in the generalized tables of standard errors may be approximated by linear interpolation.

The standard errors in tables C-1 through C-4 and the factors in table C-5 were calculated using the b parameters in table C-5. The parameters may be used directly to calculate the standard errors for estimated numbers and percentages. Methods for computation are given in the following sections.

Standard errors of estimated numbers. The approximate standard error, S_x, of an estimated number shown in this report can be obtained in two ways. It may be obtained by use of the formula

$$S_x = fs \qquad (1)$$

where f is the appropriate factor from table C-5 and s is the standard error of the estimate obtained by interpolation from table C-1 or C-2.

Alternatively, the standard error for estimates may be approximated by formula (2) from which the standard errors in tables C-1 and C-2 were calculated. Use of this formula will provide more accurate results than the use of formula (1) above.

$$S_x = \sqrt{-\frac{b}{T}x^2 + bx} \qquad (2)$$

Here x is the size of the estimate, T is the total number of persons in a specific age group and b is the parameter in table C-5 associated with the particular characteristic. If T is not known, for Total or White use 100,000,000; for Black or Hispanic use 10,000,000.

Illustration of the computation of the standard error of an estimated number. Table A-5 shows that in October 1986 there were 7,397,000 persons aged 18 to 24 years enrolled in college and 26,512,000 total persons in that age group. Using formula (1) with f = 1.0 from table C-5 and s = 110,300 from table C-1, the standard error of 7,397,000 is (1.0)(110,300) = 110,300. The value of s = 110,300 was obtained by linear interpolation in two directions in table C-1. The first interpolation was between 25,000,000 and 50,000,000 total persons for both 5,000,000 and 7,500,000 estimated persons yielding the values 96.6 and 110.9, respectively. The second interpolation was between these two values to get the value corresponding to 7,397,000 persons.

Alternatively, table C-5 indicates that the appropriate b parameter to use in calculating a standard error for this estimate is b = 2,312. Using formula (2), the approximate standard error is

$$111{,}000 \doteq \sqrt{\frac{2312}{26{,}512{,}000}(7{,}397{,}000)^2 + (2{,}312)(7{,}397{,}000)}$$

The 90-percent confidence interval is from 7,219,000 to 7,575,000 persons (using 1.6 times the standard error). Therefore, a conclusion that the average estimate from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples.

Standard errors of estimated percentages. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which this percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more. When the numerator and denominator of the percentage are in different categories, use the factor or parameter from table C-5 indicated by the numerator. The approximate standard error,

$S_{(x,p)}$, of an estimated percentage can be obtained by use of the formula

$$S_{(x,p)} = fs \tag{3}$$

In this formula, f is the appropriate factor from table C-5 and s is the standard error on the estimate from table C-3 or C-4. Alternatively, the standard error may be approximated by the following formula from which the standard errors in tables C-3 and C-4 were calculated. Use of this formula will give more accurate results than use of formula (3) above.

$$S_{(x,p)} = \sqrt{\frac{b}{x}(p)(100 - p)} \tag{4}$$

Here x is the size of the subclass of persons or households which is the base of the percentage, p is the percentage ($0 \leq p \leq 100$) is the parameter in table C-5 associated with the particular characteristic in the numerator of the percentage.

Illustration of the computation of the standard error of an estimated percentage. Table A-5 shows that in October 1986 an estimated 34.0 percent of the 21,766,000 high school graduates aged 18 to 24 years were enrolled in college. Using formula (3) with f = 1.0 from table C-5 and s = 0.5 from table C-3, the standard error of 34.0 percent is (1.0)(0.5) = 0.5. Alternatively, using formula (4) with the appropriate b parameter of 2,312 from table C-5, the standard error of 34.0 percent is given by

$$0.5 \doteq \sqrt{\frac{2,312}{21,766,000}(34.0)(66.0)}$$

Thus, a 90-percent confidence interval for this estimate, using the standard error found by formula (4), is from 33.2 to 34.8 percent.

Standard error of a difference. For a difference between two sample estimates, the standard error is approximately equal to

$$S_{(x-y)} = \sqrt{S_x^2 + S_y^2}$$

where S_x and S_y are the standard errors of the estimates x and y, respectively. The estimates can be numbers, percents, etc. This will represent the actual standard error quite accurately for the difference between two estimates of the same characteristic in two different areas or for the difference between separate and uncorrelated characteristics in the same area. If, however, there is a high positive (negative) correlation between the two characteristics, the formula will overestimate (underestimate) the true standard error.

Illustration of the calculation of the standard error of a difference. Table A-5 of this report shows that in October 1986 an estimated 83.1 percent of 22,008,000 White persons 18 to 24 years old were high school graduates as compared to 76.4 percent of 3,665,000 Black persons of the same age group. Using formula (4), the approximate standard error of 83.1 percent is 0.4, and the approximate standard error of 76.4 percent is 1.1. The apparent difference between these two estimates is 6.7 percent, and the standard error associated with the difference is

$$1.2 \doteq \sqrt{(0.4)^2 + (1.1)^2}$$

The 90-percent confidence interval on the difference of 6.7 percent is from 4.8 to 8.6 percent. Therefore, a conclusion that the average estimate of the difference derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all samples. Since this interval does not contain zero, we can conclude with 90-percent confidence that the percentage of White persons aged 18 to 24 years who were high school graduates is different from the percentage of Black persons of the same age group.

Table C-1. Generalized Standard Errors for Estimated Numbers of Persons: Total or White

(Numbers in thousands)

Estimated number of persons	Total persons in age group									
	100	250	500	1,000	2,500	5,000	10,000	25,000	50,000	100,000
10	4.6	4.7	4.8	4.8	4.8	4.8	4.8	4.8	4.8	4.8
20	6.1	6.5	6.7	6.7	6.8	6.8	6.8	6.8	6.8	6.8
30	7.0	7.8	8.1	8.2	8.3	8.3	8.3	8.3	8.3	8.3
40	7.4	8.8	9.2	9.4	9.5	9.6	9.6	9.6	9.6	9.6
50	7.6	9.6	10.2	10.5	10.6	10.7	10.7	10.7	10.7	10.7
75	6.6	11.0	12.1	12.7	13.0	13.1	13.1	13.1	13.2	13.2
100	-	11.8	13.6	14.4	14.9	15.1	15.1	15.2	15.2	15.2
200	-	9.6	16.7	19.2	20.6	21.1	21.3	21.4	21.5	21.5
300	-	-	16.7	22.0	24.7	25.5	25.9	26.2	26.3	26.3
400	-	-	13.6	23.6	27.9	29.2	29.8	30.2	30.3	30.3
500	-	-	-	24.0	30.4	32.3	33.1	33.7	33.8	33.9
750	-	-	-	20.8	34.8	38.4	40.0	41.0	41.3	41.5
1,000	-	-	-	-	37.2	43.0	45.6	47.1	47.6	47.8
2,000	-	-	-	-	30.4	52.7	60.8	65.2	66.6	67.3
3,000	-	-	-	-	-	52.7	69.7	78.1	80.7	82.0
4,000	-	-	-	-	-	43.0	74.5	88.1	92.2	94.2
5,000	-	-	-	-	-	-	76.0	96.2	102.0	104.8
7,500	-	-	-	-	-	-	65.8	110.2	121.4	126.6
10,000	-	-	-	-	-	-	-	117.8	136.0	144.2
20,000	-	-	-	-	-	-	-	96.2	166.6	192.3
30,000	-	-	-	-	-	-	-	-	166.6	220.3
40,000	-	-	-	-	-	-	-	-	136.0	235.6
50,000	-	-	-	-	-	-	-	-	-	240.4
75,000	-	-	-	-	-	-	-	-	-	208.2
100,000	-	-	-	-	-	-	-	-	-	-

Note: a. These standard errors must be multiplied by the appropriate factor in table C-5 to obtain the standard error for a specific characteristic.

b. To estimate standard errors for years 1956 to 1966 multiply the above standard errors by 1.14; for 1967 to 1980, multiply by 0.93.

c. The standard errors were calculated using the formula $\sqrt{-(b/T)\,x^2 + bx}$, where b = 2,312 (from table C-5) and T is the total number of persons in an age group.

\- Not applicable.

Table C-2. Generalized Standard Errors for Estimated Numbers of Persons: Black and Hispanic

(Numbers in thousands)

Estimated number of persons	Total persons in age group						
	100	250	500	1,000	2,500	5,000	10,000
10	4.8	5.0	5.0	5.1	5.1	5.1	5.1
20	6.4	6.9	7.1	7.1	7.2	7.2	7.2
30	7.4	8.3	8.6	8.7	8.8	8.8	8.8
40	7.9	9.3	9.8	10.0	10.1	10.2	10.2
50	8.1	10.2	10.8	11.1	11.3	11.3	11.4
75	7.0	11.7	12.9	13.4	13.8	13.9	13.9
100	-	12.5	14.4	15.3	15.8	16.0	16.0
200	-	10.2	17.7	20.4	21.9	22.3	22.6
300	-	-	17.7	23.4	26.2	27.1	27.5
400	-	-	14.4	25.0	29.6	30.9	31.6
500	-	-	-	25.5	32.2	34.2	35.1
750	-	-	-	22.1	36.9	40.7	42.5
1,000	-	-	-	-	39.5	45.6	48.4
2,000	-	-	-	-	32.2	55.9	64.5
3,000	-	-	-	-	-	55.9	73.9
4,000	-	-	-	-	-	45.6	79.0
5,000	-	-	-	-	-	-	80.6
7,500	-	-	-	-	-	-	69.8
10,000	-	-	-	-	-	-	-

Note: a. These standard errors must be multiplied by the appropriate factor in table C-5 to obtain the standard error for a specific characteristic.

b. To estimate standard errors for years 1956 to 1966 multiply the above standard errors by 1.14; for 1967 to 1980, multiply by 0.93.

c. The standard errors were calculated using the formula, $\sqrt{-(b/T)\ x^2 + bx}$, where b = 2,600 (from table C-5) and T is the total number of persons in an age group.

\- Not applicable.

Table C-3. Generalized Standard Errors for Estimated Percentages: Total or White

Base of percentage (thousands)	Estimated percentage				
	2 or 98	5 or 95	10 or 90	25 or 75	50
100	2.1	3.3	4.6	6.6	7.6
250	1.3	2.1	2.9	4.2	4.8
500	1.0	1.5	2.0	2.9	3.4
1,000	0.7	1.0	1.4	2.1	2.4
2,500	0.4	0.7	0.9	1.3	1.5
5,000	0.3	0.5	0.6	0.9	1.1
10,000	0.2	0.3	0.5	0.7	0.8
25,000	0.13	0.2	0.3	0.4	0.5
50,000	0.09	0.15	0.2	0.3	0.3
100,000	0.07	0.10	0.14	0.2	0.2
150,000	0.05	0.09	0.11	0.2	0.2

Note: a. These values must be multiplied by the appropriate factor in table C-5 to obtain the standard error for a specific characteristic.

b. To estimate standard errors for years 1956 to 1966 multiply the above standard errors by 1.14; for 1967 to 1980, multiply by 0.93.

c. The standard errors were calculated using the formula, $\sqrt{(b/x)\ p\ (100-p)}$, where b = 2,312 from table C-5.